**Project 2 Report**

**Visual Analytics**

**Group Members:**

Gulfam Ahmed Saju

Sunandini Kadiyala

Bhanu Prasad Thota

**Clustering Visualization for Named Entity Extraction in Documents**

**Abstract**

This project aims to create a web-based clustering visualization for named entity extraction in documents. The system employs the Python programming language, the Flask web framework, and the SpaCy library for natural language processing. Furthermore, the visualization is implemented using the Vis.js library. The project provides an interactive interface for users to explore and understand the relationships among various documents and their extracted named entities.

**Introduction**

Named Entity Recognition (NER) is an essential task in Natural Language Processing (NLP) that involves identifying and categorizing key information such as names of persons, organizations, locations, and dates in unstructured text. In this project, we develop a web application for visualizing the clustering of documents based on their named entities.

**Methodology**

**2.1. Clustering Documents**

For clustering documents, we use SpaCy, a popular NLP library, to perform NER on the input text files. The named entities are extracted from each document, and the extracted information is stored in a JSON format for further processing. The clustering is performed based on the presence of similar named entities in different documents.

**2.2. Design and Interaction**

The clustering visualization is designed using Vis.js, a dynamic, browser-based visualization library. The documents are represented as nodes, and the extracted named entities are also represented as nodes. The nodes are color-coded based on their types, with distinct colors for files, person names, locations, organizations, and dates. The nodes are connected with edges to show the relationships between documents and their named entities.

The visualization provides the following interactions:

Zoom in and out: Users can use the mouse wheel to zoom in and out of the visualization.

Filter by entity type: Users can interactively filter the visualization by selecting or deselecting checkboxes for each named entity type.

Navigation: Users can navigate through the visualization using the navigation buttons provided.

**2.3. Running the Code**

To run the code, follow these steps:

Install the required dependencies:

**pip install Flask spacy**

**python -m spacy download en\_core\_web\_sm**

Run the Flask application:

**python main\_tree.py**

Open a web browser and navigate to the address:

**http://localhost:5000/**

**Conclusion**

This project provides an intuitive and interactive clustering visualization for named entity extraction in documents. Users can quickly understand the relationships between documents and their named entities and filter the visualization based on the entity types of interest. This web application serves as a valuable tool for exploring and analyzing text data, offering insights into the connections among different documents.